Level of Compliance of European Union Evaluations

With United Nations Standards

Master's Thesis

Student: Hüseyin Ali Âlî Tangürek

Cornell Institute for Public Affairs

Thesis Advisor: Dr. Margaret A. Johnson

06/14/2017

Abstract

This thesis is a meta-evaluation which assesses the quality of evaluations of European Union (E.U.) financial assistance programs conducted just before and just after the announcement of a new E.U. evaluation policy known as "Better Regulation" in late 2015. In this study, evaluation standards set by the United Nations are used as the meta-evaluative criteria, while a set of recent UNICEF evaluations are used as a comparison group. Accordingly, this study uses eight evaluations from the E.U. and eight evaluations from UNICEF, with half conducted before Better Regulation and half after. To apply U.N. evaluation standards, the Global Evaluation Reports Oversight System (GEROS) of the UNICEF and its quality assessment matrix, which was designed in line with U.N. standards, is used. Results indicate that: 1) the average quality of the E.U. evaluations is much lower than that of the UNICEF evaluations and 2) the new "Better Regulation" policy framework of the E.U. has not improved the quality of E.U. evaluations so far. In the last chapter, this study proposes set of recommendations to help the E.U. improve the quality of future evaluations, and proposes a simple evaluation quality checklist with scoring reference for this purpose.

Biographical Sketch

The author of this thesis is August 2017 candidate for Master of Public Administration degree at Cornell Institute for Public Affairs at Cornell University

Huseyin Ali Ali Tangurek, a native of Ankara, Turkey, has an MPA with a concentration in social policy. Huseyin Ali Ali graduated from Uludag University in Turkey with a BA in International Relations in 2001. He has worked for Turkish Ministry of Labor and Social Security as European Union Expert from 2004 to current. He has been actively involved in dealing with Turkey-EU accession negotiations as well as performing expertise and supervision missions over the EU funded projects in the area of social policy and employment. He is married with 3 children. After the graduation he plans to come back his homeland and continue to serve his country.

Acknowledgements

I would like to express my sincere thankfulness and gratitude to Dr. Margaret Johnson, my thesis advisor, for her encouragement and guidance throughout my research. Without this direction, I would have not accomplished this work. I am grateful to her for his patiently revising and examining my thesis. I sincerely appreciate her personal and professional support during my years at Cornell.

I would like to express my deepest gratitude to my wife for her endless love, support and encouragement throughout this work.

Lastly, I would like to express my highest gratitude to great Turkish nation for her generosity to let me study in this wonderful Cornell University.

This thesis is dedicated to my beautiful wife Burcu

and my lovely children Ali Tuna, Dilcu and Ömer Alp.

Table of Contents

List of Tables

List of Figures

A. Introduction

This thesis is a meta-evaluation investigating the question: "to what extent does the Directorate-General for Regional and Urban Policy of the European Commission (DG Regio) comply with current United Nations evaluation standards (United Nations Evaluation Group, 2016) in DG Regio evaluation studies, as compared with the extent of compliance with U.N. evaluation standards by UNICEF evaluation studies?" Essentially, this study asks about the quality of recent evaluations of European Commission projects. This is an important question because the European Union's (E.U.) Regional Policy, "targets all regions and cities in the European Union in order to support job creation, business competitiveness, economic growth, sustainable development, and improve citizens' quality of life" (European Commission, 2016). The E.U. Regional Policy is managed by DG Regio, and is the E.U.'s main policy tool to improve citizen quality of life. It has a budget of € 351.8 billion, almost a third of the total E.U. budget. Therefore, evaluations conducted by DG Regio to assess the outcomes of these projects/programs are important for understanding whether the E.U. is making good use of taxpayers' money (European Commission, 2016).

Background: The cost effectiveness of E.U. interventions via its Regional Policy has been widely questioned throughout the E.U. for a long time by various parties. Meanwhile the quality of evaluations conducted by the E.U. – their methodological rigor, transparency and practical usefulness - has continued to be unknown. No meta-evaluation has yet been conducted to assess the quality of these evaluations. The lack of a common methodological framework for evaluation has continued to be a weakness for the E.U. and member states (De Peuter & De Smedt, 2006).

Methods: Considering the supranational nature of the E.U. (European Parliament, 2016), this study assumes that the United Nations (UN) is the only superior organization with whose

evaluation standards the E.U. should comply. Accordingly, to assess the quality of the evaluations conducted by the DG Regio on the regional policy programs of the E.U., this study uses the UNICEF Global Evaluation Reports Oversight System (GEROS) methodology and its assessment and rating matrix (UNICEF, 2016b) which was developed on the basis of the evaluation quality assessment standards of the U.N. (United Nations Evaluation Group, 2016). The GEROS rating matrix incorporates all principles and norms of the International Organisation of Supreme Audit Institutions (INTOSAI) as well as those of the E.U. itself, including its new Guidance Document on Monitoring and Evaluation on the European Cohesion Fund and European Regional Development Fund (European Commission, 2014), which aims to help establish a common methodology and shared understanding among E.U. institutions and member states.

The sample is composed of eight evaluations conducted by the DG Regio and eight global evaluations conducted by UNICEF (UNICEF, 2016a). Half of the sample is composed of evaluations reported in 2016, after the E.U. instituted a new evaluation policy called "Better Regulation" (European Commission, 2016a), and the other half evaluations reported before 2016. Evaluations assessing an entire program, which is managed by headquarters of DG Regio or UNICEF, aiming to make a positive difference in a specific sector (like tourism) were included in the sample. Evaluations assessing the performance of the countries, data collection studies, draft evaluations, evaluations conducted by the same contractors and other evaluations which were not officially requested or endorsed by the headquarters were excluded. Sample evaluations are assessed via GEROS assessment matrix. The study compares results for the E.U. with those for UNICEF evaluations to provide an additional frame of reference for the quality of current E.U. evaluations. It also compares evaluations reported before the E.U.'s new "Better Regulation" policy was instituted in late 2015 with those reported after the new policy was in place.

Results and Findings: The results of this study reveal that E.U. evaluation reports have a much lower rate of compliance with U.N. evaluation standards as compared with that of UNICEF evaluation reports. Further, the E.U. evaluation reports fundamentally fail to meet certain basic criteria for quality. In fact, the high end of the evaluation quality score range for E.U. reports is comparable to the low end of the evaluation quality score range for UNICEF reports. Despite the institution of the E.U.'s new "Better Regulation" evaluation policy in late 2015, the quality of the E.U. evaluation reports actually declined in 2016.

Based on these results, this study offers four fundamental findings. First, despite high expectations, the 2015 "Better Regulation" policy did not immediately increase the quality of E.U. evaluation reports. Second, E.U. officials seem to have lower expectations for the quality of evaluation reports compared with those in UNICEF. Third, E.U. evaluation contractors failed to perform high quality evaluations. Fourth, E.U. evaluations fail to provide reasonable feedback to decision makers on how to apply the results of evaluations. This may be the most critical finding of all.

Recommendations:

1) The E.U.'s "Better Regulation" policy needs to be improved to provide more useable quality control standards for evaluations.

2) E.U. evaluation staff should undergo comprehensive training programs both on general quality standards for evaluation and on the specific requirements of the "Better Regulation" policy.

3) The evaluation capacity of E.U. evaluation contractors should be substantially improved.

4) A clear policy should be developed to identify how evaluation results are to be used for improvement and decision-making relative within E.U. policies and programs.

This study ends by providing possible strategies for operationalizing the recommendations above, including a proposed simplified evaluation quality checklist and scoring reference. This tool could be used to assess the quality of future E.U. evaluations to make them more credible and useful.

B. Literature Review

This literature review will summarize existing evidence on the effectiveness of the evaluation system of the European Union in serving the interests of European citizens and regions, and explain the standards chosen in this study for assessing the quality of its recent evaluation reports. For this purpose, this chapter will summarize the main features of European Union regional policy, rules and practices on evaluating the regional policy interventions, and main evaluation standards set by the European Union. The review concludes with a summary of other authoritative evaluation standards, including those of the United Nations, which sets the stage for assessing the effects of new evaluation standards recently set forth in the European Union.

1. The European Union Regional Policy (Cohesion Policy)

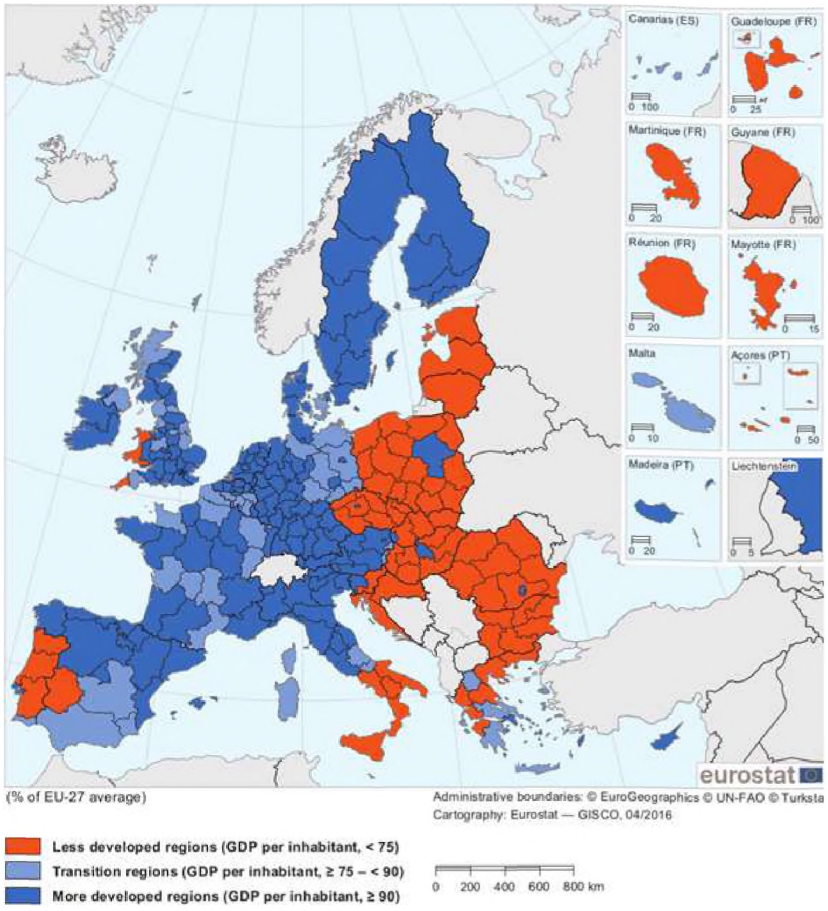
The European Union represents 509 million citizens (Worldbank, 2016) settled in 28 highly varied member states which produce $ 16,229 trillion GDP in 2015. One of the challenges of the E.U. is to deal with the considerable inequalities in income and opportunity among its different regions. The E.U. is particularly stretched by the participation of the new member states with relatively low income levels (Malais, 2009). Ensuring economic and social cohesion within the E.U. has been an objective for the E.U. since 1988, which is the date of the creation of the Regional

Policy (European Commission, 2016c). The intent of the Regional Policy is to assure redistribution of resources from rich to poor areas, narrowing down economic and social disadvantages within a system of multi-level governance (Malais, 2009).

Currently, the Regional Policy (or Cohesion Policy), which is implemented in all 276 NUTS II regions in E.U.[1], supports activities related to job creation, business competitiveness, economic growth, sustainable development, and generally improving citizens' quality of life. For the budget period of 2014-2010, it was funded at € 351.8 billion (European Commission, 2016b) out of a total E.U. budget of € 1 trillion (European Commission, 2016b). In the Cohesion Policy, more than half of the total funding (€182 million) was allocated for those regions whose GDP is lower than 75% of the average, as seen in Figure 1 below.

[1] The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU. For more information on the statistical classifications of the European Union Please see http://ec.europa.eu/eurostat/web/nuts/overview

FIGURE 1 ELIGIBILITY OF REGIONS FOR COHESION FUNDS BASED ON GDP PER INHABITANT BY NUTS II REGIONS FOR THE PERIOD OF 2014-2020.

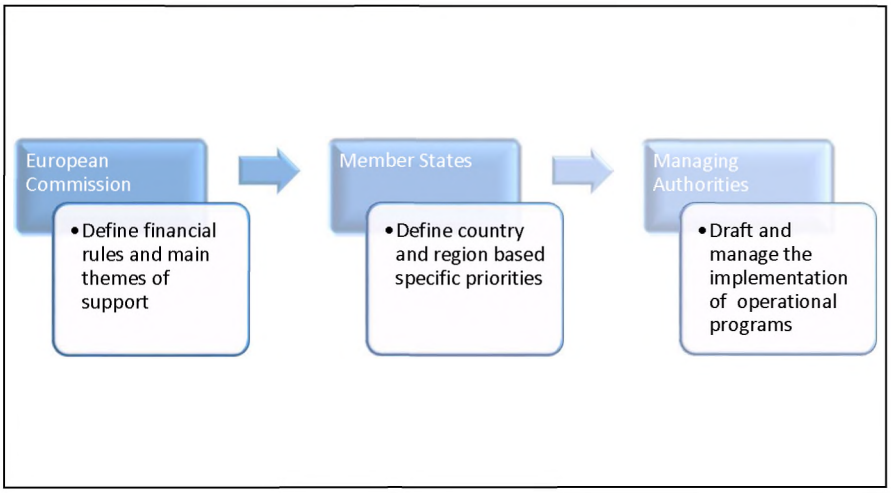


Source: European Commission, Directorate-General for Regional and Urban Policy, 2016

The Cohesion Policy has three main financial tools: 1) The European Regional Development Fund (ERDF), 2) The European Social Fund (ESF) and 3) The Cohesion Fund. These funds are mobilized through programs and projects not directly funded by the E.U.. Funds are attached to multi-annual, national programs prepared by the E.U. Member States, who co-finance the programs, in line with general E.U. objectives and priorities. These programs are discussed with the European Commission, and take final shape with the agreement of two parties. The programs are implemented in the member states by the managing authorities which are

responsible for monitoring and evaluating the program activities (Eurostat, 2016). Figure 2 below depicts the management roles in the E.U. structural funds of the European Commission, member states and their managing authorities.

FIGURE 2 MANAGEMENT STRUCTURE OF STRUCTURAL FUNDS



2. Evaluations in the European Union Regional Policy

Current guidelines for evaluation in the Regional Policy and relevant funds in the European Union are mainly based on regulations which reformed the evaluation structure in the E.U. in the late 1980s. There have been numerous changes in the regulations regarding how to conduct evaluations on projects funded through the three Cohesion funds within past 20 years (Bachtler & Wren, 2006). Currently the European Commission and member states are supposed to follow the rules stated in Regulation (E.U.) No 1303/2013 of The European Parliament and of the Council of 17 December 2013 in managing relevant E.U. funds. According to current regulation: "Evaluations shall be carried out to improve the quality of the design and implementation of programs, as well as to assess their effectiveness, efficiency and impact."

The regulation envisions distinct roles for member states and the European Commission. Three different types of evaluation must be conducted during the funding period – namely ex-ante, interim and ex-post. Ex-ante evaluations must the conducted by the relevant managing authority of a member state before the start of program implementation, to improve the program design and verify whether its objectives can be reached. At least one interim evaluation during the implementation of the program is required to assess progress toward program objectives. Accordingly, for each E.U.-funded project, the member state must draft an evaluation plan during the programming phase and submit it to the E.U.. This plan is to indicate the types and timing of evaluations which will be conducted by the managing authorities. Ex-ante and interim evaluations must subsequently be implemented in accordance with this plan. (The European Parliament and the European Council, 2013) (European Council, 2006).

Meanwhile, the European Commission assumes three different roles in terms of evaluations of the programs (apart from monitoring of the programming and implementation of the program through monitoring committees). The first role of the European Commission is related to the ex-post evaluations to examine the effectiveness and efficiency of programs after program implementation. The ex post evaluations are to be carried out by the Commission, or by the Member States in close cooperation with the Commission. The second role of the European Commission is to compile all ex-ante and ex post evaluations of the dedicated programs and submit a synthesis report to the European Parliament. This is a new role assigned by a new regulation (The European Parliament and the European Council, 2013). From this, it can be assumed that the European Commission and especially DG Regio is to take have a more active and direct role in overseeing and reporting evaluations. The third role of the European Commission is to provide guidance to member states on how to carry out evaluations. Relevant Directorates of the

Commission oversee various ways to increase evaluation capacity of the member and candidate states. They provide transition period funding, trainings, twinning activities among experts, networking events and guidance materials (Stern, 2009) which cover general information about how to design evaluation for the public administrators who are dealing with E.U. funds in member states.

Regardless of whether E.U. institutions or member states are conducting them, "evaluations shall be carried out by internal or external experts that are functionally independent of the authorities responsible for program implementation" (The European Parliament and the European Council, 2013). In line with the regulation, managing authorities in the member states or European Commission services must design the evaluation, but they should authorize these independent experts to conduct the evaluations. Following this principle, it appears evaluations performed in the member states and the European Commission are almost always conducted by private consultancy firms (European Commission DG Regio, 2016). The relevant authority designs the evaluation, and then outsources the work. Once completed, the evaluation report is submitted to the relevant authority for verification and payment.

Therefore, two crucial steps for assuring the quality of E.U. evaluations are: 1) ensuring a high-quality evaluation design while drafting the terms of reference and 2) verifying the quality of the final product when the contractor submits the final evaluation report to the relevant authority.

3. Evaluation Standards in the European Union

The European Commission was initially responsible for setting evaluation quality standards for member states in accordance with Council regulations (European Council, 2006). However, this responsibility no longer sits with the European Commission, in accordance with

more recent regulations (The European Parliament and the European Council, 2013). I could not find any specific explanation or justification for this policy change. However, it can be said that there have been clear attempts by the European Union to establish such standards.

There are two levels of studies aimed at improving the quality of the evaluations -- evaluation quality studies for member states and for the European Commission. For the member states, the most up-to-date and advanced reference is the "Guidance Document on Monitoring and Evaluation for Programming Period 2014-2020" which is a comprehensive guideline for the member states on how to perform monitoring and evaluation activities. It is a well-designed guidance document which explains what are the main features, types, timing of monitoring and evaluation activities. In terms of evaluation standards, it indicates that; "in order to ensure the quality of evaluation activities, the Commission recommends that Member States and regions base their work on clearly identified standards, established either by themselves or to use European Commission standards or those of national evaluation societies, the OECD and other organizations. Most of the standards center on general principles such as the necessity of planning, the involvement of stakeholders, transparency, use of rigorous methods and independence and dissemination of results. A summary with explanations which defines four main standards is provided in Annex 4. These four standards are:

a. *"Evaluation activities must be appropriately organized and resourced to meet their purposes.*
b. *Evaluation activities must be planned in a transparent way so that evaluation results are available in due time.*
c. *Evaluation design must provide objectives and appropriate methods and means for managing the evaluation process and its results.*

d. Evaluation activities must provide reliable and robust results."

These standards, however, fail to provide detailed evaluation quality assurance guidance for member states. Subsection 5 of the standards states that "the quality of the evaluation must be assessed on the basis of the pre-established criteria" (European Commission, 2014). In essence, the guidelines provide a list of resources which can be used by member states, a non-exhaustive list of standards as an example, and exhorts them to check the quality of the evaluations according to the job definitions determined as in the specific terms of reference for each evaluation contract.

For the European Commission, there was no binding, unified, concise, and comprehensive evaluation quality assurance policy document. Different Directorates had published different guidelines, policies or principles for the conduct of evaluation[2]. However, on May 2015 all previously published evaluation guidelines were merged into a single policy document called "Better Regulation", making possible an evaluation of the possible effects of the new regulations. The new policy has very detailed standards to ensure highest quality in impact assessment (European Commission, 2016a). It is designed to be used by Eurocracts, particularly by the people who carry evaluation responsibility in their respective Directorates. The Better Regulation policy covers all phases of the policy cycle, and provides rather complicated guidelines for users. It calls for an Internal Steering Group (ISG) to manage the evaluation process, and defines a specific

[2] Please see https://ec.europa.eu/research/evaluations/index_en.cfm?pg=home for the evaluation framework of The Directorate-General for Research and Innovation. Please see http://ec.europa.eu/europeaid/evaluation-policy_en for evaluation policy of Directorate General for International Cooperation and Development. Please see http://ec.europa.eu/regional_policy/en/policy/evaluations/ for the evaluation policy of Directorate-General for Regional and Urban Policy

document called "staff working document' (European Commission, 2016a) which is to be drafted by this group to assess the quality of the evaluation process and final product. Hence, this staff working document serves as the fundamental source for the verification of the evaluation product of the contractor. It also provides a guideline on how to follow up on the evaluation results. According to the relevant chapter of the toolbox, "the evaluation results and recommendations must feed into the Annual Activity Reports, and related follow-up actions must be identified in the Annual Management Plans of the Commission Services". It also indicates that any administrative measures can be taken immediately by the relevant E.U. institution while policy based actions must be discussed and taken by the European Commission (European Commission, 2016a). Therefore, it can be assumed that any evaluation study conducted after May 2015 could be expected to comply with the Better Regulation standards.

However, the Better Regulation guidelines do not contain a user-friendly evaluation quality assurance assessment matrix or checklist. Rather, they provide a very long and complicated list of issues to be addressed in an evaluation study. Also, the complicated procedures would appear to make it very hard to ensure the same level of information among the staff working in different Directorates. Accordingly, it is fair to be skeptical about the likelihood of positive effects from Better Regulation on the quality and use of evaluation results. The heavy complexity of Better Regulation coupled with the lack of an accessible user interface may jeopardize the best use of E.U. tax-payers' money through the structural funds.

This hypothesis can be tested if we can measure trends in E.U. evaluation quality before and after the imposition of Better Regulation. No meta-evaluation study has been conducted or published regarding the quality of evaluation reports on EC regional policy interventions. A review found only a few studies offering proposals as to how such meta-evaluations might be

conducted (Dall'erba & Fang, 2015) (Dall'erba & Fang, 2015) as well as critical narrative studies on the future shape of evaluations in the European Union (Stern, 2009).

This motivated the search for a tool through which to assess and compare the quality of evaluations before and after the 2015 Better Regulation principles were enacted.

4. International Evaluation Standards and United Nations

There are numerous evaluation standards drafted by non-governmental associations, international organizations, or individual states. None of these has been officially recognized by the international community as universal evaluation standards. Some of the most referred ones are listed below:

- **1) The United Nations Evaluation Group (UNEG) Norms and Standards for Evaluation**. Adopted in 2005 and updated in 2016, the UNEG standard "has served as a landmark document for the United Nations and beyond". It defines 86 norms under five different standards. It does not provide any assessment grid or template however determines the advanced level criteria for evaluations. (United Nations Evaluation Group, 2016);
- **2) Global Evaluation Reports Oversight System (GEROS).** Global Evaluation Reports Oversight System (GEROS) is an organization-wide system and the quality assessment tool for final evaluation reports drafted for UNICEF. It was developed based on the UNEG norms and standards. GEROS has an advanced assessment matrix for evaluating the quality of the evaluations. There are no UNICEF specific questions, however some of the questions were specifically highlighted as "highly important for UNICEF" (UNICEF, 2016b).
- **3) OECD DAC Principles for Evaluation of Development Assistance** (OECD, 1991). Developed by the OECD r to guide managers dealing with evaluations, this standard was published in 1991 and revised in 2005. It is mainly concerned with how to conduct joint evaluations by different institutions. It lacks a quality standard grid, and follows the DAC Evaluation Quality Standards (OECD, 2010) which also lacks specific assessment grids.

- **4) UNESCO Guidelines for Managing External Evaluations** (UNESCO, 2008). These guidelines describe the stages involved in managing external evaluations of which there are two broad sets: (1) those managed directly by IOS; and (2) those managed by the sector. These guidelines do not produce a new standard or detailed assessment matrix, but refer to UNEG Norms and Standards.
- **5) UNFPA Evaluation Quality Assessment** (UNFPA, 2012). Developed by UNFPA in order to be used for quality assessments by the UNFPA Evaluation Office, this standard was introduced on May 2011. It includes a basic grid which was designed as part of the efforts of UNFPA to be in line with the U.N. organizations. However, it is far less developed than the GEROS grid which is used by UNICEF.
- **6) UNDP Handbook on Planning, Monitoring and Evaluating for Development Results** (UNDP, 2011). Developed in 2009 and updated in 2011 within the framework of UNDP Evaluation Policy accepted on 2006, this handbook was designed to make the UNDP a more result-oriented institution. The handbook is dedicated to process management, and does not cover post-evaluation quality assessment process.
- 7) American Evaluation Association (AEA, 2003) / European Evaluation Society (EES, 2016). Neither of these two organizations develops or improves its own standards, but each compiles different aspects of various standards to guide evaluators, such as those of U.N. or OECD.

TABLE 1 COMPARISON TABLE FOR INTERNATIONAL EVALUATION STANDARDS

	UNEG	**OECD**	**UNICEF**	**UNESCO**	**UNFPA**	**UNDP**	**AEA/EES**
Identifies Evaluation Standards	Yes	Yes	No (Refers to UNEG)	No (Refers to UNEG)	No (Refers to UNEG)	Yes	No (Compiles Standards)
Covers Post-Evaluation Quality Assessment Standards	Yes	Yes	Refers to UNEG	Refers to UNEG	Refers to UNEG	No (Process Management)	No
Comprehensiveness of Standards	High	Medium	High	High	High	Low	No
Quality Assessment Grid/Matrix/Tool & Level	No	No	Yes Advanced	No	Yes Average	No	No

In addition to above listed evaluation standards, the International Organization of Supreme Audit Institutions (INTOSAI) has its own audit standards specifically designed for performance audits. The standards were developed in consultation with, inter alia, European Organization of Supreme Audit Institutions (EUROSAI). Those standards provide guidance to auditors regarding the fundamentals of how to conduct a performance audit. They combine the audit and auditor standards as well as process management principles. This combination of an audit-oriented approach and a lack of specific assessment grid for quality checks, makes INTOSAI standards for performance audit difficult to incorporate into this study. However, it is helpful to note that all of the main standards identified by INTOSAI are covered by U.N. Norms and Standards.

Some of the listed standards developed by international organizations like the OECD or the U.N. have had a greater opportunity to be promoted by member states. The European Union still refers to OECD standards as we have seen in the text of Guidance Document on Monitoring and Evaluation for Programming Period 2014-2020 as follows: "the Commission recommends that Member States and regions base their work on clearly identified standards, established either by themselves or to use European Commission standards or those of national evaluation societies, the OECD and other organizations" (European Commission, 2014).

On the other hand, the U.N. is the most prominent international organization in the world in which almost all countries must be a member. The United Nations Evaluation Group (UNEG) contains all the evaluation unit heads of the U.N. agencies, which consists of all nations in the world. Moreover, United Nations Norms and Standards for Evaluation, as any other similar type of document, was drafted by UNEG in cooperation with OECD Development Assistance Committee (DAC) Network on Development Evaluation (EvalNet); Evaluation Cooperation

Group (ECG); International Organisation for Cooperation in Evaluation (IOCE); and Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP).

The U.N. Norms and Standards for Evaluation seem to be the most inclusive and advanced evaluation standards ever developed in this field. On the other side, the E.U. has only recently produced its evaluation standards through the Better Regulation system. Before this, it is hard to argue that the E.U. had its own specific standards. Figures 3 and 4 below show a comparison of these two different evaluation systems. The U.N. system covers (and goes well beyond) not only the key evaluation principles of the E.U., but also its key criteria for evaluation quality. (A comparison of E.U. standards of evaluation with U.N. standards is almost impossible due to very fragmented structure of the Better Regulation guidelines and toolbox.) Based on this alone, it could be observed that the U.N. standards appear far more developed and user-friendly than the E.U. Better Regulation guidelines.

FIGURE 3 COMPARISON OF STANDARDS FOR CONDUCTING EVALUATIONS, U.N. VERSUS E.U.

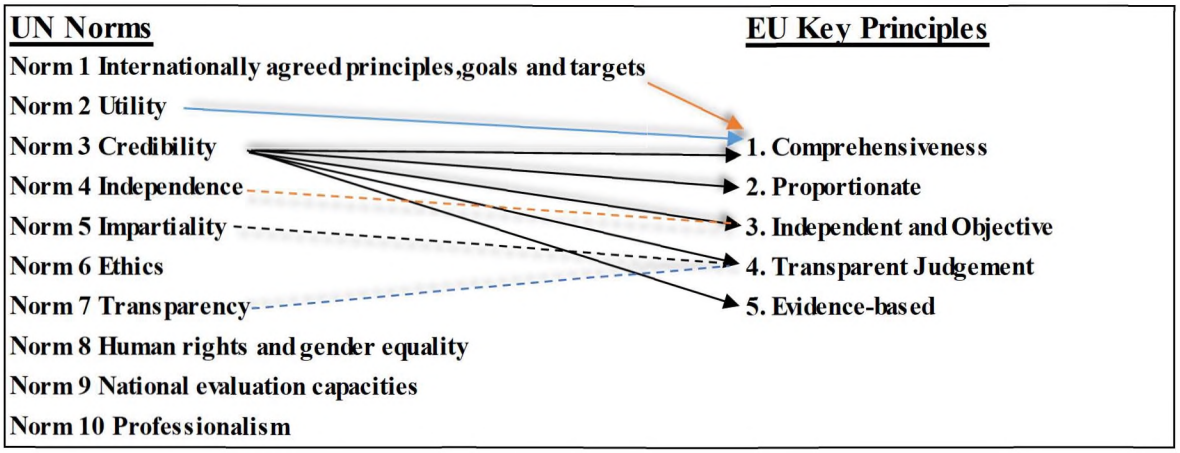
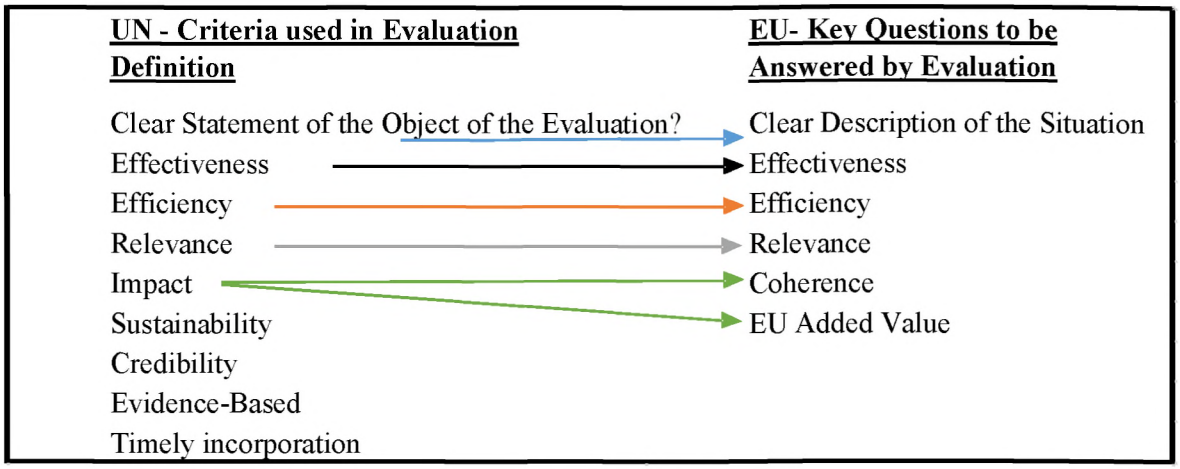


FIGURE 4 COMPARISON OF CRITERIA FOR EVALUATION QUALITY, U.N. VERSUS E.U.



It can also be argued that the U.N. is a unique supranational entity, and is thus the only superior organization to which the E.U. may bind itself for its decisions. Moreover, we can consider the U.N. Norms and Standards for Evaluation as the raw material for any evaluation quality assurance study addressing international financial assistance programs. Building on the strength of the norms and standards of the UN, UNICEF developed a meta-evaluation system called UNICEF Global Evaluation Reports Oversight System (GEROS) for use in meta-evaluation studies performed on the UNICEF funded programs, which programs were funded at $5 billion in 2015 (UNICEF, 2015). It contains a methodology and an advanced assessment and rating matrix (UNICEF, 2016b) which was developed on the basis of evaluation quality assessment standards of the U.N. (United Nations Evaluation Group, 2016). The GEROS rating matrix incorporates all principles and norms of the International Organization of Supreme Audit Institutions (INTOSAI) as well as those of the E.U. itself, including its new Guidance Document on Monitoring and Evaluation on the European Cohesion Fund and European Regional Development Fund (European Commission, 2014). It also gives cross references to OECD DAC evaluation principles and any other relevant thematic organization in case of gender, equality and ethical issues. Therefore, it is

not a closed-circuit internal use document, but a universally available methodology for assessing evaluation quality.

C. Methodology

This section explains the research methodology, including data sources, sampling methodology and limitations of the study as set forth in the design matrix used in this study, found in Annex 5. In summary, this study applies U.N. evaluation standards to rate the quality of a sample of E.U. and UNICEF evaluation reports, purposively selected to reflect the topical range of recent evaluation reports by these two institutions, both before and after the institution of the E.U.'s new "Better Regulation" policy in late 2015. Thus, the research design is pre-post evaluation with a comparison group and non-random assignment. The analysis uses a mixed-methods approach, with qualitative content analysis using a checklist, and quantitized rating and scoring.

1. Data Sources and Sampling Methodology

This thesis investigates the question: "To what extent does the Directorate-General for Regional and Urban Policy of the European Commission (DG Regio) comply with current United Nations evaluation standards (United Nations Evaluation Group, 2016) in DG Regio evaluation studies, as compared with the extent of compliance with UN evaluation standards by UNICEF evaluation studies?". It assesses E.U. evaluations in accordance with U.N. evaluation standards which will be applied through the GEROS evaluation matrix. For comparison, it applies the GEROS to check adherence levels of UNICEF evaluations, whose own evaluation policies show a high degree of alignment with those of the U.N. This provides a clear benchmark and a validity check on the GEROS matrix.

The two sources for the evaluation reports examined in this study are the UNICEF evaluation database system and the DG REGIO Publications search engine (Directorate-General for Regional and Urban Policy, 2017). This study uses a purposive sampling strategy. My sample will be composed of eight evaluations conducted by the DG Regio and eight global evaluations conducted by UNICEF. Half of the sample will be composed of the evaluations reported in 2016 and other half will be selected from the evaluations reported before 2016, since European Union did not introduce the new Better Regulation standards until late 2015. Evaluations assessing an entire program, which is managed by headquarters of DG Regio or UNICEF, aiming to make a positive difference in a specific sector (like tourism) are included in the sample. Evaluations assessing the performance of the countries, data collection studies, draft evaluations, evaluations conducted by the same contractors and other evaluations which were not officially requested or endorsed by the headquarters are excluded. Considering their proportion in the overall population of evaluation, topics of the selected samples can be considered as representative in both cases. In the case of the E.U., the selected evaluations are either main areas of intervention or they are the only applicable evaluation(s) in the selected year. The main programmatic foci for UNICEF are children and institutional responses, therefore it is fair to say that my sample of evaluation reports is representative here too. However, sample selection in the case of UNICEF was needed to select from among many global evaluations addressing the intervention of the UNICEF to crises or cases in different geographical areas. Accordingly, the UNICEF sample is representative in terms of topic but not geographic coverage, as shown in Table 2 below.

TABLE 2 COMPARISON TABLE FOR OVERALL POPULATION AND STUDY SAMPLE

	Study Topics		Geographical Locations	
	E.U.	UNICEF	E.U.	UNICEF
Overall Population	Transport, Environment, Urban and Social Infrastructure, Growth and Job Creation, Technical Assistance	Children and Institutional Responses to Crises	Entire Area of the Union	Entire World and UNICEF Headquarters
Study Sample	Transport, Environment, Urban and Social Infrastructure, Growth and Job Creation, Technical Assistance	Children and Institutional Responses to Crises	Entire Area of the Union	Syria, Central African Republic, South Sudan and Global Evaluations on Intuitional Responses of UNICEF

Sample evaluations are assessed via the GEROS assessment matrix, and a rating score assigned to each report. I then compare the average ratings for pre-2016 with those for post-2016 evaluations to obtain a "change score" for both groups. I compare average pre- and post- ratings as well as average change scores for the E.U. sample with those for the UNICEF sample. My initial expectation was that I would find quality gaps and be able to identify specific discrepancies within the E.U. sample. Within this framework, this study assesses the quality of the evaluations listed in Table 3 below.

TABLE 3 EVALUATION REPORTS IN THE STUDY SAMPLE

Year	DG REGIO	UNICEF
2016	Culture and Tourism - Final Report - Work Package 9 Ex post evaluation of Cohesion Policy programmes 2007-2013, focusing on the European Regional Development Fund (ERDF) and the Cohesion Fund (CF)	Evaluation of UNICEF's humanitarian response to the Syria crisis
	Transport - Final Report - Work Package 5 Ex post evaluation of Cohesion Policy programmes 2007-2013, focusing on the European Regional Development Fund (ERDF) and the Cohesion Fund (CF)	Evaluation of the UNICEF Response to the Crisis in the Central African Republic
	Environment - Final Report - Work package 6 Ex post evaluation of Cohesion Policy programmes 2007-2013, focusing on the European Regional Development Fund (ERDF) and the Cohesion Fund (CF)	Report of the Inter-agency Humanitarian Evaluation (IAHE) of the Response to the Crisis in South Sudan
	Urban development and Social infrastructure - Final Report - Work package 10 Ex post evaluation of Cohesion Policy programmes 2007-2013, focusing on the European Regional Development Fund (ERDF) and the Cohesion Fund (CF)	UNICEF GEROS Meta-Analysis 2015
2015	Energy efficiency in public and residential buildings - Final Report Work Package 8 - Evaluation of Cohesion Policy programmes 2007-2013	Protecting Children from Violence (VAC): A Comprehensive Evaluation of UNICEF's Strategies and Programme Performance
2013	Evaluation of the main achievements of Cohesion Policy programmes and projects over the longer term in 15 selected regions (from 1989-1993 programming period to the present)	Evaluability Assessment of the Peacebuilding, Education and Advocacy Programme (PBEA)
	Evaluation of the European Observation Network for Territorial Development and Cohesion (ESPON) programme	Evaluation of Community Management of Acute Malnutrition (CMAM): Global Synthesis Report
2012	JASPERS Evaluation	Global Evaluation of Life Skills Education Programmes

2. Data Analysis

The reports were rated using the GEROS Assessment Matrix for the content. The GEROS Matrix is a fair standard for this comparative meta-evaluation, in that it is applicable to both organizations. I generated quantitative rating grades for each question to standardize results for aggregation and comparison. (I disregarded the section of the GEROS regarding segregation of questions, since E.U. and U.N. have different standards for the evaluation document format.)

Questions and rating criteria making direct references to specific institutions, such as UNICEF, were either deleted or revised to ensure fair evaluation standards. Therefore, I have been able to avoid assessing the E.U. evaluations based on the policy priorities of the UNICEF . This applies to GEROS questions 6,12, 13, 20-24, and 55-58.

In its final version, the checklist used in this study consists of ten main questions containing 57 items in which there are two types of categories. Category 1 covers 55 questions which are asked to evaluate each specific aspect of the evaluation, while category 2 covers only two items which assess the credibility of the report as a whole within a strategic management perspective. I have rated each individual item and then calculated the total compliance score of each evaluation report by finding the percentage category 1 and category 2 items rated with either "outstanding", "yes", "mostly", "no" (which corresponded to a rating of 3, 2, 1, or 0 points respectively). No weighting system is applied to the item ratings - all questions are assumed to have equal importance. For each report, item ratings are added up and then divided by the number of main questions. The calculated result is then transformed into a percentage, e.g. 50%. Thus, the percentage adherence score for each evaluation report provides a summary statement about its quality. Under this scheme, an evaluation report could receive a high overall score in category 1 (specific aspects of quality) while receiving a substantially low rating in category 2 (overall quality in reference to a strategic framework), or vice versa.

I also prepared a scoring reference guideline which mostly follows the GEROS question rating criteria, with the exception of the revisions discussed above. For transparency, the guideline in the Annex 4 includes all questions and the respective grading scale with detailed explanations and justifications.

3. Limitations

There are four main limitations that must be taken into consideration in this meta-evaluation. As the first limitation, it is important to be aware that the Better Regulation system was introduced by the E.U. in 2015, and is thus still a very new instrument. My sample covers only four evaluation reports verified by the E.U. in 2016, just one year after the Better Regulation was issued. Therefore, results and findings regarding the Better Regulation should be considered very preliminary. Further research may be needed to fully capture possible effects of the new system.

The second limitation concerns the number of program evaluations conducted by the European Union and particularly DG REGIO. I could only find eight program evaluation studies performed by DG REGIO itself (as opposed to by E.U. member states). This situation substantially limited the sample size, and forced me to omit the year 2014.

The third limitation is the lack of official contact with the E.U. institutions. Despite several attempts to get internal information from the E.U. institutions regarding the evaluation standards, principles and practices, I could not receive any official response from anyone who might provide contextualizing background information. This of course limits the study to document review.

The last limitation pertains to the numerical precision of the qualitative results. It is not always possible to represent qualitative results with high quantitative precision. So, the quantitative results presented in this study should not be interpreted as certain numbers, but as a reasonable of indicator with moderate variance as to the overall quality of these evaluation reports. It should also be noted that having only one rater in this study is also a limitation.

D. Results and Findings

This chapter will explain the main results and findings after a modified version of the GEROS evaluation matrix was applied to the 16 evaluation reports selected from among the U.N. and UNICEF evaluation reports selected for this study.

1. Results

Result 1: Mean scores and median scores are very close in both European Commission and UNICEF samples, both for specific and general aspects of evaluation report quality.

As we see from the below in Table 3, these results show a balanced distribution in rating scores. For items on the quality of specific aspects of the evaluation report, category 1, the mean and median value in E.U. evaluations differed only 1% while it is 4% in category 2 questions. Similarly, in category 1 items for UNICEF evaluations, the difference between the mean and median values of the scores is only 1%, while it is zero for items on general evaluation report quality, category 2 items. It shows us that averages of scores in each category are meaningful, as there are no outlier scores distorting the mean.

TABLE 4 DIFFERENCE BETWEEN MEAN AND MEDIAN SCORES FOR EVALUATION REPORT QUALITY

	Specific Aspects of Evaluation Report Quality	General Evaluation Report Quality
E.U.	-1%	4%
UNICEF	-1%	0%

Result 2: Scores for specific and general aspects of evaluation quality are very close, in both European Commission and UNICEF samples.

As Table 4 indicates, there is no substantial difference among in scores for specific aspects of evaluation report quality versus scores for overall evaluation report quality, for either sample. This means that scores for the quality of specific aspects of evaluation report quality roughly correspond to scores for the overall quality of the reports.

TABLE 5 AVERAGE SCORES FOR SPECIFIC AND GENERAL ASPECTS OF EVALUATION REPORT QUALITY, E.U. V. UNICEF

	Specific Aspects of Quality	General Aspects Quality
E.U.	52%	54%
UNICEF	79%	83%

Result 3: The E.U. evaluations show a much lower level of adherence to U.N. standards than those of UNICEF.

In both categories, individual items and general aspects of quality, the evaluations contracted out by the E.U. for its own programs received scores ranging in between 50% - 54%, while UNICEF evaluations received scores ranging between 79% - 83% in the UNICEF evaluations, as seen in Table 5 below. This suggests a noticeable difference in quality of evaluation reports between the E.U. and UNICEF. The E.U. evaluation reports lack a standardized approach regarding format, criteria, methodology, results and recommendations. However, the UNICEF reports tend to follow a consistent format, and comply with certain standards such as those of OECD-DAC or the U.N. in designing and presenting methodology, results and recommendations.

Problems in the E.U. reports are not limited to adherence to U.N. standards. They include a lack of consistency in the quality of evaluation questions and methodology. In other words, each report was drafted in a different way, and each without following any clear standards for quality.

*Result 4: Scores at the **high** end of the range for E.U. reports resemble those at the **low** end of the range for UNICEF reports.*

A look at the range of scores for the E.U. and the UNICEF reports indicates that none of the E.U. reports are even close to the UNICEF's highest score reports as Table 5 shows below.

TABLE 6 COMPARISON OF EVALUATION QUALITY SCORE RANGES FOR E.U. VERSUS UNICEF EVALUATION REPORTS

	Specific Aspects of Evaluation Report Quality	General Aspects of Evaluation Report Quality
E.U.	38 — **66**%	33 -- **83**%
UNICEF	**67** -- 95%	**67** -- 100%

This result tells us a few important things. First, the quality margin between highest and lowest scores are almost equal in both categories, at around 30%. Second, it is obvious that the UNICEF evaluations scored much higher as a group compared with the set of E.U. evaluations. Third, UNICEF contractors are doing much better than the E.U. contractors in producing high quality evaluation reports.

There may be differences in how an evaluation report is approached in the two different organizations. The E.U. evaluation reports are mostly designed as the final report of a particular project, whereas the UNICEF reports are solely designed as specific evaluation reports following internationally accepted evaluation standards. A full explanation of the differences in quality found here will require further research.

Result 5: Despite the institution of a new evaluation policy raising standards, the quality of the E.U. evaluation reports actually declined after 2015.

Not only was there no increase in the quality of the evaluation reports conducted for the E.U. after the institution of Better Regulation in 2015, in fact, quality decreased. As you can see from the below Table 4x, none of the statistical measures including mean, median, lowest and highest scores under both categories show any increase in 2016. Rather, they show a perceptible decrease in the quality, despite the new evaluation quality assurance system. In fact, the lowest score for evaluation reports conducted in 2015 and before is comparable to the highest score in the evaluations conducted in 2016, in both specific quality items and overall quality.

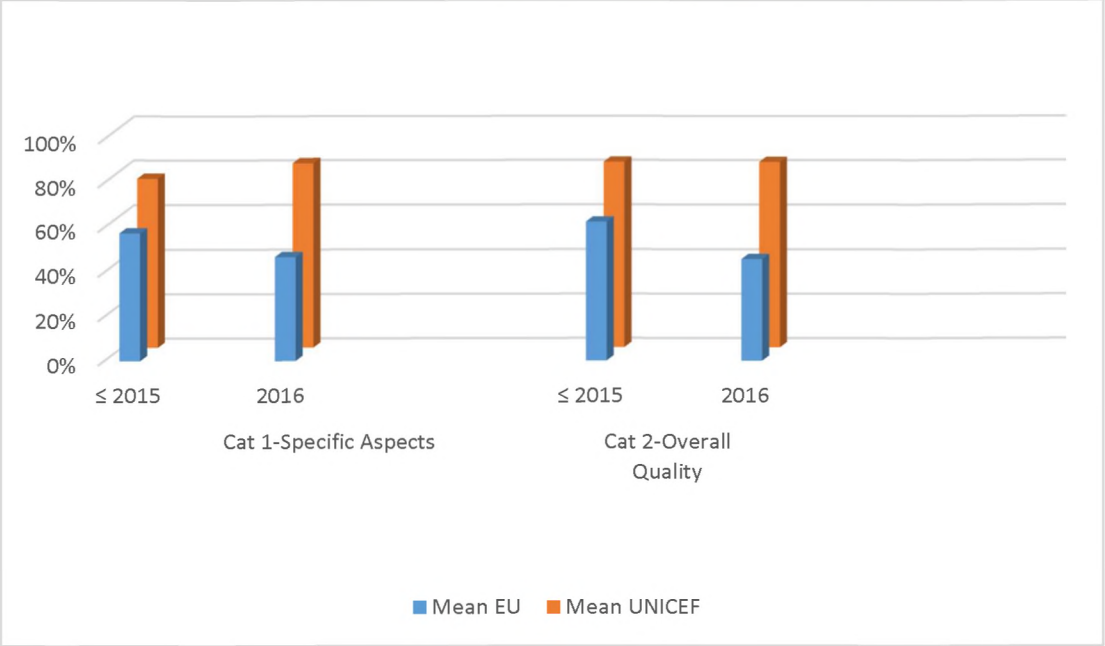
TABLE 7 COMPARISON OF EVALUATION QUALITY SCORES FOR REPORTS GENERATED BEFORE VERSUS AFTER 2015 EVALUATION POLICY CHANGE

	Average Score for Quality Pre-2015 Policy Shift	**Average Score for Quality Post-2015 Policy Shift**
E.U.	58% (range 52 – **66**)	47% (range 38 – **55**)
UNICEF	76% (range **72** – 82)	83% (range **67** – 95)

On the other hand, the UNICEF reports appear remarkably better in 2016 than previous UNICEF reports. We can see a report with 100% adherence level with the standards while the mean scores in both categories show 83 - 85% adherence levels, which is a 10-percentage point increase in the quality of the evaluations, despite the fact that there was no substantial change in the UNICEF evaluation system.

Moreover, as we see below in Figure 5, we do not see increasing quality in the E.U. evaluation reports after 2015. Rather there appears to be a slight dip in the quality of the E.U. reports. Meanwhile, the UNICEF reports show a consistent, higher level of quality over time.

FIGURE 5 EVALUATION QUALITY FOR E.U. AND UNICEF, BEFORE AND AFTER BETTER REGULATION 2015



Why didn't the new Better Regulation 2015 evaluation policy of the E.U. increase the quality of E.U. evaluation reports? One possible explanation is that the E.U. officers were not aware of the specifics of Better Regulation. If this were the case, we the quality of E.U. evaluation reports might increase in the future if more were done to raise awareness and provide training.

Another possible reason for the lack of increase in quality is that the Better Regulation policy is considerably complex, and lacks clear, step-by-step evaluation quality assessment guidelines. If this is the main reason, a simplification of the regulation and development of clear instructions may be needed to prompt an increase in the quality of E.U. evaluation reports.

It could also be that Better Regulation could not be implemented for the 2016 evaluations in this sample because these evaluation contracts were signed prior to its enactment. However, this argument could be problematic, since regardless of the ToR requirements of an evaluation contract, it is up to the contracting agency to decide its quality assurance methods. The E.U.'s heavy reliance on the ToR as a standard for quality might mean drafting better quality ToRs could

significantly increase the quality of the evaluation reports. Regardless of which check and verification process was followed by the E.U. officers, it appears a more lax standard may have been operating in 2016, rather than a stricter one.

Result 6: The E.U. evaluation reports fundamentally fail to meet certain basic criteria for quality.

In addition to the first three results of this study, I have identified some failure patterns in the E.U. reports, as seen in Figure 6 below. On the vertical axis, we see the eight E.U. evaluation reports labelled as series, and in the horizontal axis the evaluation grid questions. Score margins have been highlighted in different colors. The interesting part of this plot is the areas with no color but blue lines which indicates the density of the scores of "0" for particular questions asked to different evaluation reports. According to this, we can see that the questions "12 and 13", "20-21-22-23-24", "27" and "48" were regularly scored "0".

FIGURE 6 SCORING MAP OF THE E.U. EVALUATIONS WITHIN THE SCOPE OF THE CAT 1 - SPECIFIC QUESTIONS [3]

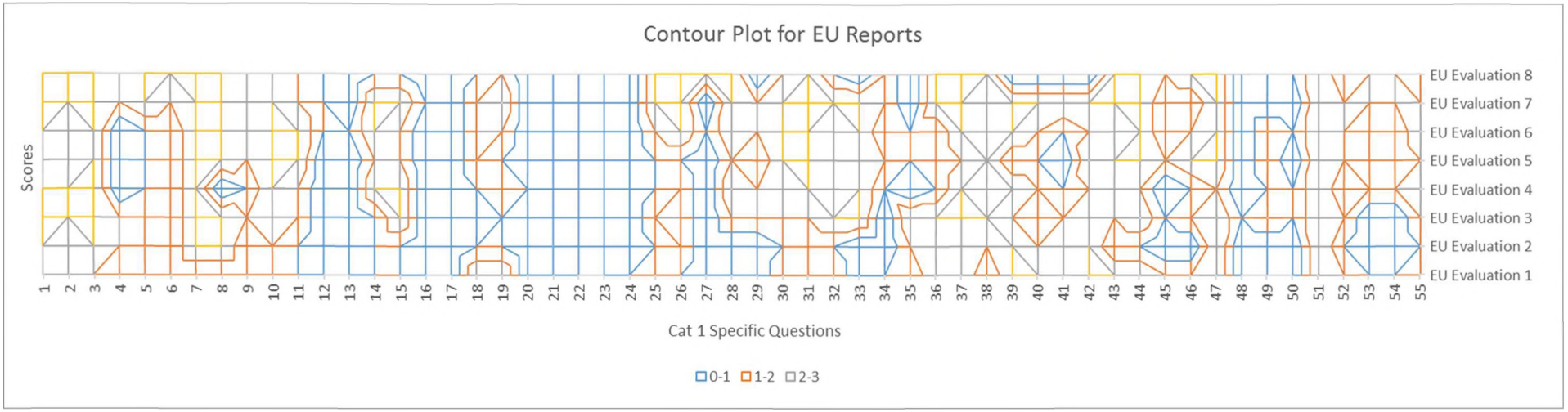


TABLE 8 FREQUENCY OF "0" SCORES RECEIVED BY THE REPORTS UNDER THE RELEVANT THEMES

Theme of the Questions	# ID of the Question	Frequency of "0" Scores in E.U. Reports (out of 8)	Frequency of "0" Scores in UNICEF Reports (out of 8
Evaluation Criteria	12-13	5 -7[4]	0-1
Ethics	16-17	7 - 8	3-5
Human Rights, Child, Gender, Equity	20-24	7-6-8-8-7	0-0-0-0-0
Use of Counterfactuals	27	6	2
Process of Developing Recommendations	48	7	4

[3] Blue areas in the figure indicate the "0" scores received by the reports from the specific questions. For instance, all the E.U. evaluations received "0" points for the 20, 21, 22, 23, 24th questions and it is indicated with frequent blue squares in the relevant area.

[4] In order of the questions in the left row

As listed in the Table 7 above, the questions scored "0" fall under certain themes, including defining and applying evaluation criteria, attending to ethical principles, considering human rights and the particular rights of children, gender equality and equity concerns, using counterfactuals and control groups, and following due process in developing recommendations. These are all universal standards for the quality of evaluations, so it is significant that there are considerable failures in these areas in the E.U. reports.

It is important to acknowledge that some aspects of the human rights-based theme, particularly children's rights, are more relevant for UNICEF work. However, human rights, gender equality and equity concerns are also important issues addressed by the E.U., as with any other international organization. The European Union should be concerned about the failure to attend to these issues in its evaluation reports.

As indicated before, scores for specific items on the GEROS evaluation quality scoring tool cannot completely account for scores for the overall quality of the reports. However, methodological problems in an evaluation can substantially affect the quality of the report if regarded through the lens of the reliability, validity and conciseness. As we see in the supportive Table 8, which articulates the lowest average scores received by 8 E.U. evaluation reports among all questions excluding the questions listed in Table 7, the E.U. evaluations fail to ensure basic quality in providing a well-articulated results chain, an acceptable statement of methodological limitations, sound analysis on cost-effectiveness, clear recommendations and appropriate and useful annexes. As a result, we can state that the E.U. evaluation reports clearly fail on 15 of the 55 quality items, apart from overall quality scores.

TABLE 9 LOWEST AVERAGE SCORES RECEIVED BY E.U. EVALUATION REPORTS

ID # of Question	Subject of the Question	E.U. Average Score (out of 3)
4	Articulation of Result Chain/Logic	1.1
29	Acceptability of Methodological Limitations	1.1
34	Discussing unexpected findings	1.1
35	Cost-analysis	1
45	Clearly Stating and Prioritizing the Recommendations	1.1
50	Wider relevance of Lessons Learned	0.5
53	Appropriateness of the Annexes	1.1
54	Usefulness and Credibility of Annexes	1.1

Table 7 above and Table 8 just above suggest that E.U. evaluation reports fail in a variety of areas. We can try to find a classification method for these items to see whether there is a pattern in these failures. To do that, I followed the method of GEROS which classifies questions under 6 main areas and found below average results for eight E.U. evaluation reports in Table 9 below.

TABLE 10 AVERAGE SCORES OF E.U. REPORTS IN DIFFERENT AREAS

Thematic Classification of Questions	Average Scores (out of 3)	Average Score Excluding Human Rights Items
Object of the Evaluation	2.2	N/A
Evaluation Purpose, Objectives and Scope	1.3	N/A
Evaluation Methodology, Gender and Human Rights	1	1.4
Findings and Conclusions	2	N/A
Recommendations and Lessons Learned	1.5	N/A
The Report is Well Structured, Logical and Clear	1.8	N/A

This summary shows that the lowest scores in the E.U. evaluations are because of the failures of the reports to define and apply proper methodology, articulate the right purpose,

objective and scope and offer well-developed recommendations and lessons learned. There is be no meaningful increase if we exclude scores for human rights, children rights, women rights and equity related questions. The E.U. evaluations score better for explaining the object of the evaluation and listing findings and conclusions. Scores for structure, logic and clarity of the reports are lower.

2. Findings

Finding 1: Better Regulation, the new evaluation policy of the E.U., did not immediately increase the quality of the E.U. reports

There is a substantial difference in quality of the UNICEF and E.U. evaluation reports in this sample. E.U. evaluation reports, in general, do not appear to comply with any pre-defined evaluation quality standards except for the specific Terms of Reference for each report. That said, it is not always possible to find the Terms of Reference in the annex of the reports. It is not within the scope of this study to check the compliance level of the reports to the Terms of References (ToR). However, the findings in this study strongly suggest either the quality of ToRs in this sample is low, or monitoring and verification mechanisms in the E.U. headquarters are weak. At a minimum, we can say that there are no evaluation quality standards systematically followed by the E.U. officials in either assuring that ToRs are included in evaluation reports, or in verifying the quality of these reports.

It is fair to say that the Better Regulation evaluation policy change failed to have an immediate positive effect on the quality of the evaluation reports. This is to be expected, to some extent, since the smart regulation directs the E.U. officials to use only the ToRs for checking the quality of any evaluation reports, and the ToRs for my sample of evaluation reports had been

drafted long before the better regulation was issued. On the other hand, it is obvious that there needs to be a mechanism for controlling the quality of ongoing evaluations, since it may be years before we see evaluation reports prepared in accordance with ToRs which comply with Better Regulation. However, it is unlikely the quality of ToRs will increase in the future unless the administrators of Better Regulation provide a new, user-friendly and exhaustive evaluation quality assurance assessment matrix or checklist.

Finding 2: E.U. officials seem to have lower expectations for the quality of evaluation reports.

This study showed scores at the high end of the range for E.U. reports fall just above those at the low end of the range for UNICEF report. Bearing in mind that all of the sample reports were verified by the DG Regio, the level of expectations for quality of an evaluation report during the period studied appears to have been significantly lower in the E.U. than in UNICEF.

There may be a few reasons behind this. First, E.U. officers may view evaluation reports as they do any other project final report, without considering evaluation standards. Therefore, their way of regarding the quality of any project final report (merely: did it comply with the ToR?) may guide their approach verifying evaluation reports. This may point to a broader problem of low standards for quality and effectiveness of the E.U. Structural Funds in general. Another reason may be a general lack of awareness among the E.U. officials as to evaluation quality standards and their importance. Better Regulation was instituted quite recently, however this does not explain the lower quality of the reports verified before 2016 as compared with those of UNICEF, and the apparent decrease in the quality of E.U. evaluation reports verified in 2016. There appears to be a pressing need to improve the evaluation capacity of E.U. staff overseeing grant funded projects.

Last, but not least, we can say that it is very hard to ensure the same level of information among the staff working in different the E.U. Directorates due to the complicated procedures of Better Regulation. Its excessive complexity and abstractness makes it difficult for staff to understand how to implement it, let alone train others in how to do so.

Finding 3: The E.U. evaluation contractors failed to perform high quality evaluations.

There is a substantial difference in evaluation quality between the E.U. and UNICEF evaluations. One of the most important factors in the system is the competence of those who perform the evaluation activities. Apart from the fact that they mostly follow only the ToR requirements, and the fact that there are design weaknesses in these ToRs, the E.U. contractors in this study failed to understand the evaluation objective, purpose and scope and to address them with a sound evaluation methodology. Regardless of the quality of the ToR, the E.U. contractors in this sample also failed to generate useful recommendations which might have had a positive effect on future E.U. policies. These failures cannot be fully explained by ambiguities in Better Regulation or the monitoring and verification weaknesses of E.U. officials. So why?

It could be that the procurement procedures in the E.U. place contractors under pressure to decrease their costs at the expense of the quality. Accordingly, quality may substantially decrease due to low compensation to evaluators coupled with evaluators' need to keep the profit margin reasonable. One another possible explanation for the low quality of evaluation contractor outputs might be the lack of a clear understanding of the specific demands of any given evaluation. Lack of internal expertise on evaluation in the contractor firms may force them to seek external experts who cannot be well supervised by the firms for proper quality assurance. Thus, there may be need for capacity building in contractor firms and freelance evaluators as to appropriate methods and other aspects of quality in the evaluations they generate.

Finding 4: The E.U. evaluations fail to provide reasonable feedback to decision makers.

We need to consider two different implications of this finding. The first is that the low quality of these evaluations makes it harder to draw any clear conclusions about the effectiveness and efficiency of programs supported by the E.U. Structural Funds. A second is that the lack of clear and reasonable recommendations in these evaluation reports means the E.U. is funding evaluations yet failing to obtain from them what is needed to improve the efficiency and effectiveness of its programs.

These two insights call into question why these evaluation studies are being conducted. Low-quality, vague evaluations serve to reduce the level of accountability of public officials, since the quality of officials' performance in managing these funds is not being appropriately measured, and is therefore unknown. Having uninformative evaluations may create an artificial atmosphere of high quality of outputs in E.U. headquarters, undermining any possibility of action to improve the system. Similarly, irrelevant or trivial recommendations may create another artificial atmosphere in which E.U. officials think they are improving the system, but actually are not.

E. Recommendations

In this section, I conclude with some recommendations to improve the quality of the evaluations conducted by the E.U.. These recommendations will be based on the findings explained in the latter section and listed in order of importance.

Recommendation 1: Better Regulation needs to be improved so as to provide more workable quality control standards for evaluations.

It was found in this study that the Better Regulation has so far failed to improve the quality of E.U. evaluations, perhaps due to its complicated, fragmented and ambiguous quality control standards.

For proper implementation, Better Regulation needs a well-designed quality control checklist. This checklist should be integrated into evaluation ToR documents in order to let the evaluation contractors know what will be the standard for quality of the evaluation study. This checklist should be addressed by the contractors in each reporting phase and used by the monitoring and verification staff of the E.U. Directorates to ensure consistency. As part of these recommendations, I drafted a sample checklist which was prepared in line with the latest GEROS rating matrix approach, found in Annex 1. This checklist is a simplified version of the matrix used to assess the quality of evaluations in this study. Most parts of this recommended checklist are from the simplified GEROS matrix used by the UNICEF. The reason this simplified version is recommended for the E.U. evaluations is to assure ease of use by E.U. staff who are also checking ToR requirements. To further facilitate use, a scoring reference table is provided in Annex 2.

Better coordination among the different E.U. Directorates is necessary to ensure consistent implementation of quality standards in evaluation studies. It is recommended that an Evaluation Coordination Unit be established to secure coordination in drafting evaluation ToRs, verifying evaluation reports and following up on evaluation results.

Recommendation 2: Evaluation staff must participate in comprehensive training programs both on basic quality standards for evaluation and on the specific principles of the Better Regulation.

Better Regulation covers all of the fundamental aspects of an evaluation even though it does not have a specific and easy to use checklist for the quality assessment. Moreover, the E.U. has longstanding evaluation practices conducted by different Directorates. However, the low evaluation quality standards of the evaluation reports show a remarkable need for changing and improving the preconditions in the minds of the evaluation staff on basic standards for evaluation reports.

To tackle this, evaluation staff in different Directorates should participate in ongoing training programs. Such programs should have three objectives. The first objective should be to change the ideas about what constitutes quality in evaluations. Within this objective, the staff should be made familiar with the theoretical principles of program evaluation and best practices in the scope of international evaluation practices. The second objective should be to introduce the Better Regulation principles, including a practical checklist, to the relevant staff. The third objective should be to harmonize the evaluation quality approaches of the different Directorates and their staff. Those objectives must be realized with both theoretical and practical trainings. On the job trainings should be organized and external expertise should be mobilized to increase the qualifications of the E.U. staff. Inter-organizational cooperation may work well to ensure an overall increase in evaluation standards. UNICEF may be a very good partner for this purpose.

The decision makers should be aware of the fact that the low quality standards for evaluations may be pointing to low standards in all jobs contracted out by the E.U.. Therefore, even though it is out of the scope of this study, it may be useful for the E.U. to conduct a meta-evaluation study, perhaps by an experienced UNICEF evaluation contractor, to evaluate the effectiveness and efficiency of the E.U. financial assistance system in general.

Recommendation 3: Capacity of the evaluation contractors must be substantially improved.

There appear to be a number of high capacity evaluation contractors working on E.U. evaluation contracts. If so, it seems that this capacity has not been fully engaged to perform high quality evaluations. It may be that the lack of an E.U. evaluation quality assessment framework or low expectations of E.U. staff have negatively affected the performance of these contractors. However, it is possible the selection of evaluation contractors may be another determinant of evaluation quality.

Assuming that there are plenty of good evaluators and contractors in the market, the E.U. ToR drafters and decision makers should revisit their budgeting and expert specification practices. While this study did not examine the amount of evaluation contract awards, it is possible that unbalanced budget allocations or moderate expert specifications for evaluation contracts may negatively affect the quality of the evaluators nominated by the contractors, and as a result, the evaluation method and implementation of the evaluation contractors. The E.U. should not compromise from the quality of the expertise when drafting the evaluation ToRs.

Should there be a limited capacity of the E.U. contractors in employing or reaching out to good evaluators, then the E.U. should revisit its contracting out strategy so as to attract more advanced contractors and evaluators to its contracts.

Regardless of the current capacity of the contractors, the E.U. should conduct and awareness raising campaign among the contractors to make them aware of Better Regulation standards.

Recommendation 4: A policy framework should be developed guide the systematic use of evaluation results.

The effectiveness of an intervention is often measured by its level of success in achieving predefined objectives. This approach is valid for projects and programs as well as evaluations. Similarly, evaluations can be designed to assess the effectiveness, efficiency, preparedness of a program, project or policy. However, the effectiveness of an evaluation lies also in the extent to which the results are used for improvement and decision making.

Given the criticisms of the effectiveness and efficiency the European Union financial assistance programs and these findings on the low quality of E.U. evaluations, the E.U. would also be well served by developing a policy framework which identifies how evaluation results will be used to improve the effectiveness and efficiency of the financial assistance programs. This policy framework should have very clear instructions for evaluated subjects on how to apply evaluation recommendations. It should also give authority to a unit (such as an Evaluation Coordination Unit) to conduct regular follow up studies to check the progress on the adoption of evaluation recommendations.

F. Bibliography

AEA. (2003). Guiding Principles for Evaluators. Retrieved December 16, 2016, from http://www.eval.org/p/cm/ld/fid=51

Bachtler, J., & Wren, C. (2006). Evaluation of European Union Cohesion policy: Research questions and policy challenges. *Regional Studies*, *40*(2), 143–153. https://doi.org/10.1080/00343400600600454

Dall'erba, S., & Fang, F. (2015). Meta-Analysis of the Impact of European Union Structural Funds on Regional Growth. *Regional Studies*, 1–12.

De Peuter, B., & De Smedt, J. (2006). Problems of Policy Evaluation in a European Multi-Level Governance Context: The case of Active Labour Policy in Belgium. UKES/EES Joint International Conference Evaluation in Society: Critical Connections, London, UK.

Directorate-General for Regional and Urban Policy. (2017). Publications. Retrieved June 14, 2017, from http://ec.europa.eu/regional_policy/en/information/publications?title=&themeId=0&typeId=4&countryId=0&periodId=0&fundId=0&policyId=0&languageCode=en

EES. (2016). International and Supranational Organisations. Retrieved December 16, 2016, from http://europeanevaluation.org/resources/evaluation-standards/international-and-supranational-organisations

European Commission. (2014, March). Guidance Document on Monitoring and Evaluation for Programming Period 2014-2020; European Cohesion Fund and European Regional Development Fund. Retrieved from http://ec.europa.eu/regional_policy/sources/docoffic/2014/working/wd_2014_en.pdf

European Commission. (2016a). Better Regulation - Guidelines on evaluation and Fitness Checks - European Commission. Retrieved November 2, 2016, from http://ec.europa.eu/smart-regulation/guidelines/ug_chap6_en.htm

European Commission. (2016b). Multiannual Financial Framework-European Commission. Retrieved October 26, 2016, from http://ec.europa.eu/budget/mff/index_en.cfm

European Commission. (2016c). The E.U.'s main investment policy. Retrieved September 30, 2016, from http://ec.europa.eu/regional_policy/en/policy/what/investment-policy/

European Commission. (2016d). Tool #47: The Staff Working Document for Evaluation. Retrieved December 15, 2016, from http://ec.europa.eu/smart-regulation/guidelines/tool_47_en.htm

European Commission. (2016e). Tool #49: Follow Up Action Plans. Retrieved December 15, 2016, from http://ec.europa.eu/smart-regulation/guidelines/tool_49_en.htm

European Commission DG Regio. (2016). Evaluations by the Member States. Retrieved October 31, 2016, from http://ec.europa.eu/regional_policy/en/policy/evaluations/member-states/

European Parliament. (2016, June). Supranational decision-making procedures | E.U. fact sheets | European Parliament. Retrieved September 30, 2016, from http://www.europarl.europa.eu/atyourservice/en/displayFtu.html?ftuId=FTU_1.4.1.html

Eurostat. (2016). Regional policies and Europe 2020 - Statistics Explained. Retrieved October 26, 2016, from http://ec.europa.eu/eurostat/statistics-explained/index.php/Regional_policies_and_Europe_2020#What_is_cohesion_policy.3F

Malais, J. (2009). European Union Regional Policy. *School of Doctoral Studies European Union*, 77.

OECD. (1991). The DAC Principles for the Evaluation of Development Assistance. Retrieved from http://www.oecd.org/dac/evaluation/2755284.pdf

OECD. (2010). DAC Evaluation Quality Standards. Retrieved from http://www.oecd.org/development/evaluation/qualitystandards.pdf

Stern, E. (2009). Evaluation policy in the European Union and its institutions. *New Directions for Evaluation*, *2009*(123), 67–85. https://doi.org/10.1002/ev.306

The European Parliament and the European Council. Regulation (E.U.) No 1303/2013, Pub. L. No. 1303/2013 (2013). Retrieved from http://eur-lex.europa.eu/legal-content/en/ALL/?uri=celex:32013R1303

UNDP. (2011). The Handbook on Planning, Monitoring and Evaluating for Development Results. Retrieved December 16, 2016, from http://web.undp.org/evaluation/guidance.shtml#handbook

UNESCO. (2008). Guidelines for Managing External Evaluations. Retrieved from http://unesdoc.unesco.org/images/0015/001583/158395E.pdf

UNFPA. (2012). 2012 Evaluation Quality Assessment. Retrieved December 16, 2016, from http://www.unfpa.org/admin-resource/2012-evaluation-quality-assessment

UNICEF. (2015). *Annual Report 2015*. Retrieved from https://www.unicef.org/publications/files/UNICEF_Annual_Report_2015_En.pdf

UNICEF. (2016a). 2016 | Evaluation database. Retrieved October 12, 2016, from http://www.unicef.org/evaldatabase/index_91122.html

UNICEF. (2016b). Global Evaluation Reports Oversight System (GEROS). Retrieved October 6, 2016, from http://www.unicef.org/evaluation/index_GEROS.html

United Nations Evaluation Group. (2016). *Detail of Norms and Standards for Evaluation (2016)*. United Nations. Retrieved from http://www.unevaluation.org/document/download/2601

Worldbank. (2016). European Union | Data. Retrieved October 26, 2016, from http://data.worldbank.org/region/european-union

G. Annexes

Annex 1: Recommended Simple Checklist for Use With EU Evaluations

SAMPLE CHECKLIST FOR EVALUATIONS					
IDENTITY INFORMATION					
Title of the Evaluation					
ID NO:	1				
Year of Publication	Authoring Org	Sponsoring Org	Type of Evaluation	Comments	
CHECKLIST					
Section 1	BACKGROUND				COMMENTS
1	Is the object of the evaluation clearly described?			Grading Score Here	
2	Is the context of the intervention clearly described?				
3	Is the results chain or logic well articulated?				
4	Are key stakeholders and their contributions clearly identified				
Section 2	EVALUATION PURPOSE, OBJECTIVES AND SCOPE				
5	Is the purpose of the evaluation clearly described?				
6	Are the objectives and scope of the evaluation clear and realistic?				
7	Does the evaluation provide a relevant list of evaluation criteria that are explicitly justified as appropriate for the purpose of the evaluation?				
8	Does the report specify methods for data collection, analysis, and sampling?				
9	Are ethical issues and considerations described?				
Section 3	EVALUATION FINDINGS				
10	Do the findings clearly address all evaluation objectives and scope?				
11	Are evaluation findings derived from the conscientious, explicit and judicious use of the best available, objective, reliable and valid data.				
12	Are evaluation findings derived by accurate quantitative and qualitative analysis of evidence.				
13	Does the evaluation assess and use the intervention's Monitoring system?				
Section 4	EVALUATION CONCLUSIONS & LESSONS LEARNED				
14	Do the conclusions present an objective overall assessment of the intervention?				
15	Are lessons learned correctly identified?				
Section 5	RECOMMENDATIONS				
16	Are recommendations well grounded in the evaluation?				
17	Are recommendations clearly presented?				
Overall	EVALUATION STRUCTURE/PRESENTATION				
18	Does the evaluation report include all relevant information?				
19	Is the report logically structured?				
Rating of the Evaluation					

Annex 2 Recommended Scoring Reference Table for Use With the Simple Checklist

Scoring Reference Table			
No = 0	Mostly= 1	Yes=2	Outstanding=3
QUESTION AND CRITERIA			
1 Is the object of the evaluation well described? - Clear and relevant description of the intervention, including: location(s), timelines, cost/budget, and implementation status - Clear and relevant description of intended beneficiaries by type (i.e., institutions/organisations; communities; individuals...), by geographic location(s) (i.e., urban, rural, particular neighbourhoods, town/cites, sub-regions...) and in terms of numbers reached (as appropriate to the purpose of the evaluation) - Description of the relative importance of the object to UNICEF (e.g. in terms of size, influence, or positioning)			
2 Is the context explained and related to the object that is to be evaluated? - Clear and relevant description of the context of the intervention (policy, socio-economic, political, institutional, international factors relevant to the implementation of the intervention) - Clear and relevant description (where appropriate) of the status and needs of the target groups for the intervention - Explanation of how the context relates to the implementation of the intervention			
3 Is the results chain or logic well-articulated? - Clear and complete description of the intervention's intended results - Intervention logic presented as a coherent theory of change, logic chain or logic framework			
4 Are key stakeholders clearly identified? - Identification of implementing agency(ies), development partners, primary duty bearers, secondary duty bearers, and rights holders - Identification of the specific contributions and roles of key stakeholders (financial or otherwise), including the E.U.			
5 Is the purpose of the evaluation clearly described? - Specific identification of how the evaluation is intended to be used and to what this use is expected to achieve - Identification of appropriate primary intended users of the evaluation			
6 Are the objectives and scope of the evaluation clear and realistic? - Clear and complete description of what the evaluation seeks to achieve by the end of the process with reference to any changes made to the objectives included in the ToR - Clear and relevant description of the scope of the evaluation: what will and will not be covered (thematically, chronologically, geographically with key terms defined), as well as the reasons for this scope (e.g., specifications by the TORs, lack of access to particular geographic areas for political or safety reasons at the time of the evaluation, lack of data/evidence on particular elements of the intervention)			

7 Does the evaluation provide a relevant list of evaluation criteria that are explicitly justified as appropriate for the purpose of the evaluation? (The E.U. evaluation standards refer to the OECD/DAC criteria. Not all OECD/DAC criteria are relevant to all evaluation objectives and scopes. Standard OECD DAC Criteria include: Relevance; Effectiveness; Efficiency; Sustainability; Impact. Evaluations should also consider equity, gender and human rights (these can be mainstreamed into other criteria). - Clear and relevant presentation of the evaluation framework including clear evaluation questions used to guide the evaluation - If the framework is other than the E.U. standard criteria, or if not all standard criteria of the chosen framework are included, the reasons for this are clearly explained and the chosen framework is clearly described
8 Does the report specify methods for data collection, analysis, and sampling? - Clear and complete description of a relevant design and set of methods that are suitable for the evaluation's purpose, objectives and scope - Clear and complete description 0f the data sources, rationale for their selection and sampling strategy. This should include a description of how diverse perspectives are captured (or if not, provide reasons for this), how accuracy is ensured, and the extent to which data limitations are mitigated - Clear and complete description of the methods of analysis, including triangulation of multiple lines and levels of evidence (if relevant)? - Clear and complete description of limitations and constraint
9 Are ethical issues and considerations described? (The evaluation can be guided by the UNEG ethical standards for evaluation. As such, the evaluation report should include:) - Explicit reference to the obligations of evaluators (independence, impartiality, credibility, conflicts of interest, accountability) - Description of ethical safeguards for participants appropriate for the issues described (respect for dignity and diversity, right to self-determination, fair representation, compliance with codes for vulnerable groups, confidentiality, and avoidance of harm)
10 Do the findings clearly address all evaluation objectives and scope? - Findings marshal sufficient levels of evidence to systematically address all of the evaluation's questions and criteria - If feasible and relevant to the purpose, cost analysis is clearly presented (how costs compare to similar interventions or standards, most efficient way to get expected results)-if not feasible, an explanation is provided - Reference to the intervention's results framework in the formulation of the findings
11 Are evaluation findings derived from the conscientious, explicit and judicious use of the best available, objective, reliable and valid data - The evaluation clearly presents multiple lines (including multiple time series) and levels (output, outcome, and appropriate disaggregation) of credible evidence. - Findings are clearly supported by and respond to the evidence presented, including both positive and negative. Findings are based on clear performance indicators, standards, benchmarks, or other means of comparison - Unexpected effects (positive and negative) are identified and analysed
12 Are evaluation findings derived by accurate quantitative and qualitative analysis of evidence. - The causal factors (contextual, organisational, managerial, etc.) leading to achievement or non-achievement of results are clearly identified. For theory-based evaluations, findings analyse the logical chain (progression -or not- from implementation to results).

13 Does the evaluation assess and use the intervention's Monitoring system? - Clear and comprehensive assessment of the intervention's monitoring system (including completeness and appropriateness of results/performance framework -including vertical and horizontal logic; M&E tools and their usage) - Clear and complete assessment of the use of monitoring data in decision making
14 Do the conclusions present an objective overall assessment of the intervention? - Clear and complete description of the strengths and weaknesses of the intervention that adds insight and analysis beyond the findings - Description of the foreseeable implications of the findings for the future of the intervention (if formative evaluation or if the implementation is expected to continue or have additional phase) - The conclusions are derived appropriately from findings
15 Are lessons learned correctly identified? - Correctly identified lessons that stem logically from the findings, presents an analysis of how they can be applied to different contexts and/or different sectors, and takes into account evidential limitations such as generalizing from single point observations.
16 Are recommendations well-grounded in the evaluation? - Recommendations are logically derived from the findings and/or conclusions - Recommendations are useful to primary intended users and uses (relevant to the intervention and provide realistic description of how they can be made operational in the context of the evaluation) - Clear description of the process for developing recommendations, including a relevant explanation if the level of participation of stakeholders at this stage is not in proportion with the level of participation in the intervention and/or in the conduct of the evaluation
17 Are recommendations clearly presented? - Clear identification of target group for action for each recommendation (or clearly clustered group of recommendations) - Clear prioritization and/or classification of recommendations to support use
18 Does the evaluation report include all relevant information? - Opening pages include: Name of evaluated object, timeframe of the evaluation, date of report, location of evaluated object, names and/or organization(s) of the evaluator(s), name of organization commissioning the evaluation, table of contents -including, as relevant, tables, graphs, figures, annexes-; list of acronyms/abbreviations, page numbers - Annexes should include, when not present in the body of the report: Terms of Reference, Evaluation matrix, list of interviewees, list of site visits, data collection instruments (such as survey or interview questionnaires), list of documentary evidence. Other appropriate annexes could include: additional details on methodology, copy of the results chain, information about the evaluator(s)
19 Is the report logically structured? - The structure is easy to identify and navigate (for instance, with numbered sections, clear titles and sub-titles - Context, purpose and methodology would normally precede findings, which would normally be followed by conclusions, lessons learned and recommendations

Annex 3: Evaluation Quality Checklist Used in This Study

IDENTITY INFORMATION				
Title of the Evaluation				
ID NO:	1			
Year of Publication	**Authoring Org**	**Sponsoring Org**	**Type of Evaluation**	**Comments**

ID of Questions	CHECKLIST (with GEROS Reference Question Numbers)		COMMENTS
1	Is the object of the evaluation well described?		
2	Is the context explained and related to the object that is to be evaluated?		
3	Does this illuminate findings?		
4	Is the results chain or logic well-articulated?		
5	Are key stakeholders clearly identified?		
6	Are key stakeholders' contributions described?		
7	Are UNICEF/E.U. contributions described?		
8	Is the implementation status described?		
9	Is the purpose of the evaluation clear?		
10	Are the objectives and scope of the evaluation clear and realistic?		
11	Do the objective and scope relate to the purpose?		
12	Does the evaluation provide a relevant list of evaluation criteria that are explicitly justified as appropriate for the Purpose?		
13	Does the evaluation explain why the evaluation criteria were chosen and/or any standard evaluation criteria (above) rejected?		
14	Does the report specify data collection methods, analysis methods, sampling methods and benchmarks?		
15	Does the report specify data sources, the rationale for their selection, and their limitations?		
16	Are ethical issues and considerations described?		

17	Does the report refer to ethical safeguards appropriate for the issues described?		
18	Is the capability and robustness of the evaluated object's monitoring system adequately assessed?		
19	Does the evaluation make appropriate use of the M&E framework of the evaluated object?		
20	Did the evaluation design and style consider incorporation of the U.N. and UNICEF's commitment to a human rights-based approach to programming, to gender equality, and to equity?		
21	Does the evaluation assess the extent to which the implementation of the evaluated object was monitored through human rights (inc. gender & child rights) frameworks?		
22	Do the methodology, analytical framework, findings, conclusions, recommendations & lessons provide appropriate information on HUMAN RIGHTS (inc. women & child rights)?		
23	Do the methodology, analytical framework, findings, conclusions, recommendations & lessons provide appropriate information on GENDER EQUALITY AND WOMEN'S EMPOWERMENT?		
24	Do the methodology, analytical framework, findings, conclusions, recommendations & lessons provide appropriate information on EQUITY?		
25	Are the levels and activities of stakeholder consultation described?		
26	Are the levels of participation appropriate for the task in hand?		
27	Is there an attempt to construct a counterfactual or address issues of contribution/attribution?		
28	Does the methodology answer the evaluation questions in the context of the evaluation?		
29	Are methodological limitations acceptable for the task in hand?		
30	Are findings clearly presented and based on the objective use of the reported evidence?		

31	Do the findings address all of the evaluation's stated criteria and questions?		
32	Do findings demonstrate the progression to results based on the evidence reported?		
33	Are gaps and limitations discussed?		
34	Are unexpected findings discussed?		
35	Is a cost analysis presented that is well grounded in the findings reported?		
36	Does the evaluation make a fair and reasonable attempt to assign contribution for results to identified stakeholders?		
37	Are causal reasons for accomplishments and failures identified as much as possible?		
38	Are the future implications of continuing constraints discussed?		
39	Do the conclusions present both the strengths and weaknesses of the evaluated object?		
40	Do the conclusions represent actual insights into important issues that add value to the findings?		
41	Do conclusions take due account of the views of a diverse cross-section of stakeholders?		
42	Are the conclusions pitched at a level that is relevant to the end users of the evaluation?		
43	Are the recommendations well-grounded in the evidence and conclusions reported?		
44	Are recommendations relevant to the object and the purpose of the evaluation?		
45	Are recommendations clearly stated and prioritised?		
46	Does each recommendation clearly identify the target group for action?		
47	Are the recommendations realistic in the context of the evaluation?		
48	Does the report describe the process followed in developing the recommendations?		
49	Are lessons learned correctly identified?		
50	Are lessons learned generalised to indicate what wider relevance they may have?		

51	Do the opening pages contain all the basic elements?		
52	Is the report logically structured?		
53	Do the annexes contain appropriate elements?		
54	Do the annexes increase the usefulness and credibility of the report?		
55	Is an executive summary included as part of the report? Does the executive summary contain all the necessary elements? Can the executive summary stand alone? Can the executive summary inform decision making?		
Category 1 Rating of the Evaluation		0%	
	Category 2 Questions		
i	To what extent does each of the six sections of the evaluation provide sufficient credibility to give the reasonable person confidence to act?		
ii	To what extent do the six sections hold together in a logically consistent way that provides common threads throughout the report?		
Category 2 Rating of the Evaluation		0%	

Annex 4: Scoring Reference Table Used in This Study

Scoring Reference Table			
No = 0	Mostly= 1	Yes=2	Outstanding=3
QUESTION AND CRITERIA			
Category 1 Questions			
1 Is the object of the evaluation well described? This needs to include a clear description of the interventions (project, programme, policies, otherwise) to be evaluated including how the designer thought that it would address the problem identified, implementing modalities, other parameters including costs, relative importance in the organization and (number of) people reached.			
2 Is the context explained and related to the object that is to be evaluated? The context includes factors that have a direct bearing on the object of the evaluation: social, political, economic, demographic, and institutional. These factors may include strategies, policies, goals, frameworks & priorities at the: international level; national Government level; individual agency level			
3 Does this illuminate findings? The context should ideally be linked to the findings so that it is clear how the wider situation may have influenced the outcomes observed.			
4 Is the results chain or logic well articulated? The report should identify how the designers of the evaluated object thought that it would address the problem that they had identified. This can include a results chain or other logic models such as theory of change. It can include inputs, outputs and outcomes, it may also include impacts. The models need to be clearly described and explained.			
5 Are key stakeholders clearly identified? These include o implementing agency(ies) - development partners rights holders - primary duty bearers - secondary duty bearers			
6 Are key stakeholders' contributions described? This can involve financial or other contributions and should be specific.			
7 Are UNICEF/E.U. contributions described? This can involve financial or other contributions and should be specific			
8 Is the implementation status described? This includes the phase of implementation and significant changes that have happened to plans, strategies, performance frameworks, etc. that have occurred - including the implications of these changes.			
9 Is the purpose of the evaluation clear? This includes why the evaluation is needed at this time, who needs the information, what information is needed, how the information will be used.			
10 Are the objectives and scope of the evaluation clear and realistic? This includes: - Objectives should be clear and explain what the evaluation is seeking to achieve; - Scope should clearly describe and justify what the evaluation will and will not cover; Evaluation questions may optionally be included to add additional details			

11 Do the objective and scope relate to the purpose?
The reasons for holding the evaluation at this time in the project cycle (purpose) should link logically with the specific objectives the evaluation seeks to achieve and the boundaries chosen for the evaluation (scope)

12 Does the evaluation provide a relevant list of evaluation criteria that are explicitly justified as appropriate for the Purpose?
It is imperative to make the basis of the value judgements used in the evaluation transparent if it is to be understood and convincing. UNEG evaluation standards refer to the OECD/DAC criteria, but other criteria can be used such as Human rights and humanitarian criteria and standards (e.g. SPHERE Standards) but this needs justification. Not all OECD/DAC criteria are relevant to all evaluation objectives and scopes. The TOR may set the criteria to be used, but these should be (re)confirmed by the evaluator. Standard OECD DAC Criteria include: Relevance; Effectiveness; Efficiency; Sustainability; Impact (Specific E.U. criteria counts)

13 Does the evaluation explain why the evaluation criteria were chosen and/or any standard evaluation criteria (above) rejected?
The rationale for using each particular criterion and rejecting any standard OECD-DAC /E.U. criteria (where they would be applicable) should be explained in the report.

14 Does the report specify data collection methods, analysis methods, sampling methods and benchmarks?
This should include the rationale for selecting methods and their limitations based on commonly accepted best practice.

15 Does the report specify data sources, the rationale for their selection, and their limitations?
This should include a discussion of how the mix of data sources was used to obtain a diversity of perspectives, ensure accuracy & overcome data limits

16 Are ethical issues and considerations described?
The design of the evaluation should contemplate: How ethical the initial design of the programme was; The balance of costs and benefits to participants (including possible negative impact) in the programme and in the evaluation; The ethics of who is included and excluded in the evaluation and how this is done

17 Does the report refer to ethical safeguards appropriate for the issues described?
When the topic of an evaluation is contentious, there is a heightened need to protect those participating. These should be guided by the UNICEF Evaluation Office Technical Note and include: protection of confidentiality; protection of rights; protection of dignity and welfare of people (especially children); Informed consent; Feedback to participants; Mechanisms for shaping the behaviour of evaluators and data collectors

18 Is the capability and robustness of the evaluated object's monitoring system adequately assessed?
The evaluation should consider the details and overall functioning of the management system in relation to results: from the M&E system design, through individual tools, to the use of data in management decision making

19 Does the evaluation make appropriate use of the M&E framework of the evaluated object?
In addition to articulating the logic model (results chain) used by the programme, the evaluation should make use of the object's logframe or other results framework to guide the assessment. The results framework indicates how the programme design team expected to assess effectiveness, and it forms the guiding structure for the management of implementation.

20 This could be done in a variety of ways including: use of a rights-based framework, use of CRC, CCC, CEDAW and other rights related benchmarks, analysis of right holders and duty bearers and focus on aspects of equity, social exclusion and gender. Style includes: using human-rights language; gender-sensitive and child sensitive writing; disaggregating data by gender, age and disability groups; disaggregating data by socially exluded groups.

21 UNICEF commits to go beyond monitoring the achievement of desirable outcomes, and to ensure that these are achieved through morally acceptable processes. The evaluation should consider whether the programme was managed and adjusted according to human rights and gender monitoring of processes.

22 The inclusion of human rights frameworks in the evaluation methodology should continue to cascade down the evaluation report and be obvious in the data analysis, findings, conclusions, any recommendations and any lessons learned. If identified in the scope the methodology should be capable of assessing the level of: Identification of the human rights claims of rights-holders and the corresponding human rights obligations of duty-bearers, as well as the immediate underlying & structural causes of the non realisation of rights.; Capacity development of rights-holders to claim rights, and duty-bearers to fulfil obligations.

23 The inclusion of gender equality frameworks in the evaluation methodology should continue to cascade down the evaluation report and be obvious in the data analysis, findings, conclusions, any recommendations and any lessons learned. If identified in the scope the methodology should be capable of assessing the immediate underlying & structural causes of social exclusion; and capacity development of women to claim rights, and duty-bearers to fulfill their equality obligations.

24 The inclusion of equity considerations in the evaluation methodology should continue to cascade down the evaluation report and be obvious in the data analysis, findings, conclusions, any recommendations and any lessons learned. If identified in the scope the methodology should be capable of assessing the capacity development of rights-holders to claim rights, and duty-bearers to fulfill obligations & aspects of equity.

25 Are the levels and activities of stakeholder consultation described?
This goes beyond just using stakeholders as sources of information and includes the degree of participation in the evaluation itself. The report should include the rationale for selecting this level of participation. Roles for participation might include:
- Liaison
- Technical advisory
- Observer
- Active decision making

The reviewer should look for the soundness of the description and rationale for the degree of participation rather than the level of participation itself.

26 Are the levels of participation appropriate for the task in hand?
The breadth & degree of stakeholder participation feasible in evaluation activities will depend partly on the kind of participation achieved in the evaluated object. The reviewer should note here whether a higher degree of participation may have been feasible & preferable.

27 Is there an attempt to construct a counterfactual or address issues of contribution/attribution?
The counterfactual can be constructed in several ways which can be more or less rigorous. It can be done by contacting eligible beneficiaries that were not reached by the programme, or a theoretical counterfactual based on historical trends, or it can also be a comparison group.

28 Does the methodology answer the evaluation questions in the context of the evaluation?
The methodology should link back to the Purpose and be capable of providing answers to the evaluation questions.

29 Are methodological limitations acceptable for the task in hand?
Limitations must be specifically recognised and appropriate efforts taken to control bias. This includes the use of triangulation, and the use of robust data collection tools (interview protocols, observation tools etc). Bias limitations can be addressed in three main areas: Bias inherent in the sources of data; Bias introduced through the methods of data collection; Bias that colours the interpretation of findings

30 Are findings clearly presented and based on the objective use of the reported evidence?
Findings regarding the inputs for the completion of activities or process achievements should be distinguished clearly from results. Findings on results should clearly distinguish outputs, outcomes and impacts (where appropriate). Findings must demonstrate full marshalling and objective use of the evidence generated by the evaluation data collection. Findings should also tell the 'whole story' of the evidence and avoid bias.

31 Do the findings address all of the evaluation's stated criteria and questions?
The findings should seek to systematically address all of the evaluation questions according to the evaluation framework articulated in the report.

32 Do findings demonstrate the progression to results based on the evidence reported?
There should be a logical chain developed by the findings, which shows the progression (or lack of) from implementation to results.

33 Are gaps and limitations discussed?
The data may be inadequate to answer all the evaluation questions as satisfactorily as intended, in this case the limitations should be clearly presented and discussed. Caveats should be included to guide the reader on how to interpret the findings. Any gaps in the programme or unintended effects should also be addressed

34 Are unexpected findings discussed?
If the data reveals (or suggests) unusual or unexpected issues, these should be highlighted and discussed in terms of their implications.

35 Is a cost analysis presented that is well grounded in the findings reported?
Cost analysis is not always feasible or appropriate. If this is the case then the reasons should be explained. Otherwise the evaluation should use an appropriate scope and methodology of cost analysis to answer the following questions:
- How programme costs compare to other similar programmes or standards
- Most efficient way to get expected results
- Cost implications of scaling up or down
- Cost implications for replicating in a different context
- Is the programme worth doing from a cost perspective
- Costs and the sustainability of the programme

36 Does the evaluation make a fair and reasonable attempt to assign contribution for results to identified stakeholders?
For results attributed to the programme, the result should be mapped as accurately as possible to the inputs of different stakeholders.

37 Are causal reasons for accomplishments and failures identified as much as possible?
These should be concise and usable. They should be based on the evidence and be theoretically robust.

38 Are the future implications of continuing constraints discussed?
The implications can be, for example, in terms of the cost of the programme, ability to deliver results, reputational risk, and breach of human rights obligations.

39 Do the conclusions present both the strengths and weaknesses of the evaluated object?
Conclusions should give a balanced view of both the stronger aspects and weaker aspects of the evaluated object with reference to the evaluation criteria and human rights based approach.

40 Do the conclusions represent actual insights into important issues that add value to the findings?
Conclusions should go beyond findings and identify important underlying problems and/or priority issues. Simple conclusions that are already well known do not add value and should be avoided.

41 Do conclusions take due account of the views of a diverse cross-section of stakeholders?
As well as being logically derived from findings, conclusions should seek to represent the range of views encountered in the evaluation, and not simply reflect the bias of the individual evaluator. Carrying these diverse views through to the presentation of conclusions (considered here) is only possible if the methodology has gathered and analyzed information from a broad range of stakeholders.

42 Are the conclusions pitched at a level that is relevant to the end users of the evaluation?
Conclusions should speak to the evaluation participants, stakeholders and users. These may cover a wide range of groups and conclusions should thus be stated clearly and accessibly: adding value and understanding to the report (for example, some stakeholders may not understand the methodology or findings, but the conclusions should clarify what these findings mean to them in the context of the programme).

43 Are the recommendations well-grounded in the evidence and conclusions reported?
Recommendations should be logically based in findings and conclusions of the report.

44 Are recommendations relevant to the object and the purpose of the evaluation? Recommendations should be relevant to the evaluated object
45 Are recommendations clearly stated and prioritised? If the recommendations are few in number (up to 5) then this can also be considered to be prioritised. Recommendations that are over-specific or represent a long list of items are not of as much value to managers. Where there is a long list of recommendations, the most important should be ordered in priority.
46 Does each recommendation clearly identify the target group for action? Recommendations should provide clear and relevant suggestions for action linked to the stakeholders who might put that recommendation into action. This ensures that the evaluators have a good understanding of the programme dynamics and that recommendations are realistic.
47 Are the recommendations realistic in the context of the evaluation? This includes: * an understanding of the commissioning organisation * awareness of the implementation constraints * an understanding of the follow-up processes
48 Does the report describe the process followed in developing the recommendations? The preparation of recommendations needs to suit the evaluation process. Participation by stakeholders in the development of recommendations is strongly encouraged to increase ownership and utility.
49 Are lessons learned correctly identified? Lessons learned are contributions to general knowledge. They may refine or add to commonly accepted understanding, but should not be merely a repetition of common knowledge. Findings and conclusions specific to the evaluated object are not lessons learned.
50 Are lessons learned generalised to indicate what wider relevance they may have? Correctly identified lessons learned should include an analysis of how they can be applied to contexts and situations outside of the evaluated object.
51 Do the opening pages contain all the basic elements? Basic elements include all of: Name of the evaluated object; Timeframe of the evaluation and date of the report; Locations of the evaluated object; Names and/or organisations of evaluators; Name of the organisation commissioning the evaluation; Table of contents including tables, graphs, figures and annex; List of acronyms
52 Is the report logically structured? Context, purpose, methodology and findings logically structured. Findings would normally come before conclusions, recommendations & lessons learnt
53 Do the annexes contain appropriate elements? Appropriate elements may include: ToRs; List of interviewees and site visits; List of documentary evidence; Details on methodology; Data collection instruments; Information about the evaluators; Copy of the evaluation matrix; Copy of the Results chain. Where they add value to the report
54 Do the annexes increase the usefulness and credibility of the report?
55 Is an executive summary included as part of the report? Does the executive summary contain all the necessary elements? Can the executive summary stand alone? Can the executive summary inform decision making? Necessary elements include all of: Overview of the evaluated object; Evaluation objectives and intended audience; Evaluation methodology; Most important findings and conclusions; Main recommendations It should not require reference to the rest of the report documents and should not introduce new information or arguments. It should be short (ideally 2-3 pages), and increase the utility for decision makers by highlight key priorities.
Category 2 Questions **Informed by the answers above, apply the reasonable person test to answer the following question: Ω/ Is this a credible report that addresses the evaluation purpose and objectives based on evidence, and that can therefore be used with confidence? This question should be considered from the perspective of strategic management.**

i. Taken on their own, could a reasonable person have confidence in each of the evaluation elements separately? It is particularly important to consider: o Is the report methodologically appropriate? o Is the evidence sufficient, robust and authoritative? o Do the analysis, findings, conclusions and recommendations hold together?
ii. The report should hold together not just as individually appropriately elements, but as a consistent and logical 'whole'.

Annex 5: Design Matrix Used in This Study

Evaluation Question	**Scope & Methodology**			**Strengths & Limitations**	**Possible Claims**
	Sampling & Data	Design	Analysis		
To what extent does the Directorate-General for Regional and Urban Policy of the European Commission comply with current United Nations evaluation standards (United Nations Evaluation Group, 2016) in DG Regio evaluation studies, as compared with the extent of compliance with U.N. evaluation standards by UNICEF evaluation studies?	Purposive sample of 8 evaluations conducted by the DG Regio (group 1) and 8 global evaluations conducted by UNICEF (group 2) . Within each group, half were reported in 2016 and half reported before 2016[5]	Pre/post with comparison group and non-random assignment 1 N O X O 2 N O O where X is the imposition of E.U. Better Regulation evaluation standards in late 2015 and where O is data collection	Apply the content analysis checklist of the GEROS rating matrix. Compute an overall score for each evaluation. Compute descriptive statistics. Compare mean and range of scores for E.U. and UNICEF evaluations prior to and during 2016	Strengths: Purposive selection of samples allows for meaningful comparisons. Previous professional experience of the author can help to contextualize the situation. Limitations: Early phase in implementation of E.U. Better Regulation standards makes findings preliminary at best. Limited number of program evaluations conducted by the E.U. Lack of official contact with E.U. institutions	The E.U. evaluation reports in this sample adhere to U.N. standards by ... percent , with an average score of ... The quality of E.U. evaluations in this sample is the same as/lower than/higher than those of UNICEF. The quality of evaluations reported during or after 2016 is the same/lower than/ higher than of evaluations reported before 2016.

[5] Evaluations assessing an entire program, which is managed by headquarters of DG Regio or UNICEF, aiming to make a positive difference in a specific sector (like tourism) will be included in the sample. Evaluations assessing the performance of the countries, data collection studies, draft evaluations, evaluations conducted by the same contractors and other evaluations which were not officially requested or endorsed by the headquarters will be excluded.